Integrating High Order Ambisonics and Deep Learning for Advanced Instrument Separation in Spatial Audio Applications

Jaime Garcia-Martinez, Pablo Cabanas-Molero, Pedro Vera-Candeas,
Julio J. Carabias-Orti, Antonio J. Munoz-Montoro
Telecommunication Engineering Department
Universidad de Jaén, Spain

***Abstract*—This work explores the integration of Higher-Order Ambisonics (HOA) and deep learning for advanced sound source separation in spatial audio applications. We present a computationally efficient spherical harmonics (SH) beamforming framework that extracts spatial components from raw microphone array recordings. Our methodology leverages SH-based spatial filtering combined with a deep learning de-bleeding model, enhancing source separation in audio applications. We evaluate the performance of the SH beamformer and neural network model both independently and in combination, demonstrating that the proposed pipeline achieves superior source isolation while maintaining high signal fidelity. Unlike previous studies that relied on synthetic data, our approach is validated with real-world recordings captured using a third-order spherical microphone array. Results highlight the effectiveness of integrating spatial domain filtering with deep learning for reducing interference and enhancing separation quality. Furthermore, we provide an open-source implementation of our approach to encourage its adoption in spatial audio processing tasks, including music production. Our findings pave the way for a portable-studio paradigm, relying solely on an HOA array.**

***Index Terms*—Ambisonics, Source Separation, Spatial Audio, Deep Learning, Music Production, Real-world Recordings.**

I. Introduction

Spatial audio technologies have gained significant attention in recent years, enabling immersive auditory experiences in applications such as virtual reality (VR), augmented reality (AR), music production, and teleconferencing. Among these, Ambisonics has emerged as a widely used method for capturing and reproducing three-dimensional sound fields. Ambisonics represents the recorded sound scenario using Spherical Harmonics (SH) decomposition, providing a mathematically elegant and flexible framework for spatial sound encoding [1].

Ambisonic recording setups rely on spherical microphone arrays that capture sound from all directions, projecting recordings onto the SH domain. This enables spatial sound manipulation through beamforming and signal processing techniques. Regardless of the spatial audio method, the goal remains to synthesize virtual microphones suitable for reproduction or further processing.

Previous works, such as [2], have explored generating virtual microphones with adjustable directivity patterns through filtering operations, rather than using the SH framework. While effective, this approach is inherently tied to specific microphone array configurations. In contrast, the SH domain enables flexible beamforming techniques [3], allowing virtual microphones to be computed independently of the microphone array geometry.

Recent deep learning advances have significantly improved sound source separation [4]. By leveraging large-scale datasets, neural networks have surpassed traditional methods in speech enhancement, music separation, and spatial audio processing. Works combining deep learning with spatial filtering [5] demonstrate how spatial cues can enhance source separation based on direction of arrival (DOA) and spatial distribution.

While prior studies such as [6]–[8] primarily used synthetic Ambisonic signals, we focus on real-world microphone array recordings, addressing practical challenges like signal-to-noise ratio (SNR) degradation due to excessive amplification at low frequencies in higher-order SH components.

The main contributions of this paper are as follows:

- An SH separation framework integrating an SH beamformer with a neural network-based de-bleeding approach, enabling advanced instrument separation in spatial audio.
- An open-source implementation[1] for converting spherical microphone array recordings into SH components, featuring an efficient axis-symmetric SH beamforming algorithm. This release also includes the measured room impulse responses (RIRs) used in this work, providing a valuable resource for further research in Ambisonic processing and source separation.
- A demonstration of the practical applicability of the method through real-world microphone array recordings, expanding its relevance beyond synthetic datasets and proving its effectiveness in real acoustic environments.

This work was supported by “REPERTORIUM” Project under Grant Agreement 101095065. Horizon Europe. Cluster II. Culture, Creativity and Inclusive society. Call HORIZON-CL2-2022-HERITAGE-01-02.

[1] The proposed implementation is available at https://github.com/QHPC-SP-Research-Lab/SHArrayBeamforming.

Here, we evaluate both the SH beamformer and a neural network model as standalone separation methods, establishing baselines to assess the benefits of their combined use. By integrating spatial filtering with deep learning, our approach facilitates portable-studio music production, significantly reducing the need for multiple microphones and simplifying ensemble and band recordings with spherical microphone arrays.

II. Spherical harmonics domain pressure field representation

The complex spherical harmonics [9] are mathematically defined as follows:

$$Y_n^m(\theta, \phi) = \sqrt{\frac{2n+1}{4\pi}\frac{(n-m)!}{(n+m)!}} P_n^m(cos(\theta))e^{im\phi} \tag{1}$$

where n and m denote the order and degree of the spherical harmonics, respectively, $P_n^m(.)$ is the associated Legendre function, and $i = \sqrt{-1}$.

The pressure field can be expanded in terms of the spherical harmonics as:

$$p(\kappa, r, \theta, \phi) \approx \sum_{n=0}^{N} \sum_{m=-n}^{n} p_{nm}(\kappa, r) Y_n^m(\theta, \phi) \tag{2}$$

with the Spherical Fourier Transform coefficients denoted as $p_{nm}(\kappa, r)$. Note that the maximum order N of analysis in the Spherical Fourier Transform domain is determined by the total number of microphones Q of the array. Specifically, the minimum number of microphones Q required for the Spherical Fourier analysis is $(N+1)^2$.

Equation (3) is known as plane-wave decomposition [10], and expresses the sound pressure field coefficients in the Spherical Fourier Transform domain $p_{nm}(\kappa, r)$ in terms of a function describing the radial dependence of the field $b_n(\kappa, r)$ and the plane-wave amplitude density $\alpha_{nm}(\kappa)$:

$$p_{nm}(\kappa, r) = b_n(\kappa r)\alpha_{nm}(\kappa) \tag{3}$$

where $\kappa = \frac{2\pi f}{c}$ is the wavenumber, c is the speed of sound in air and f denotes frequency. By computing the coefficients $\alpha_{nm}(\kappa)$, a spherical harmonic representation for the pressure field that is independent of the array geometry is obtained [11].

When the microphones of the array are mounted on a spherical rigid surface, the function $b_n(\kappa r)$ is expressed as:

$$b_n(\kappa r) = 4\pi i^n \left[j_n(\kappa r) - \frac{j_n^{'}(\kappa r_a)}{h_n^{(2)'}(\kappa r_a)} h_n^{(2)}(\kappa r) \right] \tag{4}$$

where $j_n(\cdot)$ is the spherical Bessel function of the first kind and order n, $h_n^{(2)}(\cdot)$ is the spherical Hankel function of the second kind and order n, r_a is the radius of array sphere and the $'$ operator denotes the first derivative of the corresponding function. The magnitude of $b_n(\kappa r)$ is represented in Figure 1 for $r = r_a$ and up to order $N = 3$.

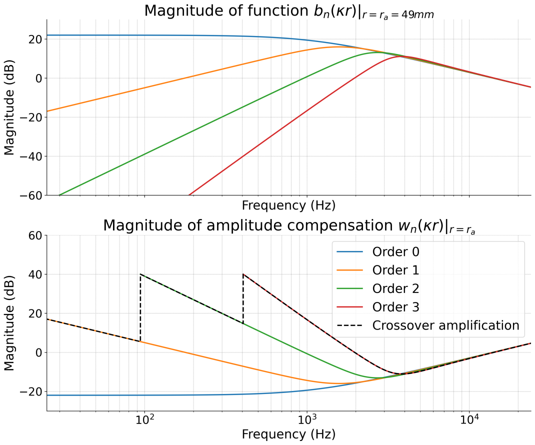

Fig. 1. Magnitude of the radial function of Equation 4 evaluated for frequencies ranging from 20 to 24000 Hz (top) and multiplicative inverse of the same functions (bottom). The crossover frequencies are selected to ensure an amplification lower than 40 dB when computing $\frac{p_{nm}(\kappa,r)}{b_n(\kappa r)}$.

A. Spherical harmonics beamformer

The general expression of an axis-symmetric beamformer in the Spherical Fourier Transform domain [10] is given by:

$$x_v(\kappa) = \sum_{n=0}^{N} \sum_{m=-n}^{n} \frac{p_{nm}(\kappa, r)}{b_n(\kappa r)} d_n(\kappa) Y_n^m(\theta_v, \phi_v) \tag{5}$$

where the pointing direction of the beamformer is expressed in spherical coordinates (θ_v, ϕ_v) and $d_n(\kappa)$ are the beamformer coefficients that define an axis-symmetric polar pattern with desired directivity properties, which may depend on the frequency. The $d_n(\kappa)$ coefficients, when applied to the SH transform of the pressure field, shape the resulting pressure field to include the desired filter properties. In this work, a beamformer pointing in a specific spatial direction will be referred to as a “virtual microphone” for the remainder of the paper.

III. Proposed SH separation pipeline

The proposed SH separation pipeline introduces a spatial-informed, neural network-based de-bleeding approach for processing microphone array signals. Figure 2 shows the schematic representation of the proposed pipeline. This pipeline takes as input both the microphone array signals and direction of arrival information, which defines the pointing directions of the virtual microphones produced by the SH beamformer implemented in this work.

Considering a spherical microphone array of Q microphones, the recorded signal $x_q[m]$ by each microphone $q \in \{0, 1, ..., Q-1\}$ is a combination of contributions from S distinct sound sources. Each source $s \in \{0, 1, ..., S-1\}$ emits a signal $s_s[m]$ that propagates to the microphones through a reverberant and noisy environment. Thus, the total sound pressure field recorded at microphone q is represented as:

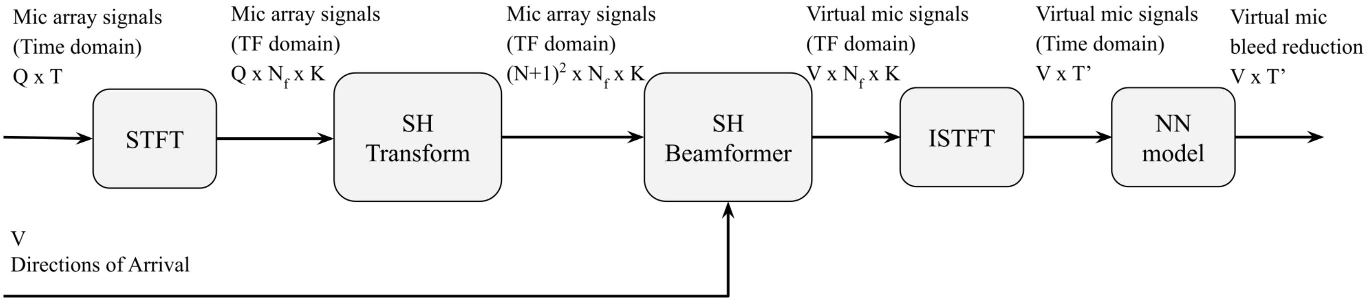

Fig. 2. Schematic representation of the proposed SH separation pipeline. The process begins with Q microphone array signals in the time domain, which are transformed into the time-frequency (TF) domain. The Spherical Harmonics transform is then applied, enabling the computation of an SH beamformer steered towards the input V directions of arrival (one per sound source in our experimentation). This results in the TF representation of virtual microphone signals, which are subsequently converted back to the time domain. Finally, a neural network-based bleed reduction stage refines each virtual microphone signal independently, producing the final processed output. Notably, the dimensionality of the signals progressively decreases as they pass through the pipeline.

$$x_q[m] = \sum_{s} s_s[m] * h_{q,s}[m] + n_q[m] \tag{6}$$

where $*$ denotes the linear convolution operator, $h_{q,s}[m]$ is the room impulse response between source s and microphone q, capturing the effects of reverberation and sound propagation delay and $n_q[n]$ is the noise captured by microphone q, which includes background noise and sensor noise.

Given that the microphone array samples the sound pressure field on the surface of a sphere, the recorded signals correspond to a discrete spatial and temporal sampling of this field over the microphone array's surface. Moreover, the spherical array considered in this work consists of a single rigid sphere with all microphones mounted on its surface, which implies that $r_q = r_a \forall q$, with r_a denoting the radius of the array sphere.

A. SH beamformer implementation

As depicted in Figure 2, the SH beamformer used in this work operates in the time-frequency domain, processing the Short-Time Fourier Transform (STFT) [12] representations $X_q[n_f, k]$ of the recorded microphone signals $x_q[m]$ for microphone q, with n_f and k indexing time frames and frequency bins, respectively.

At each STFT time frame n_f, the beamforming operation in (5) can be formulated in matrix form as presented in (7). Note that, instead of directly working with κ, we use the discrete frequency index k, which corresponds to a linear frequency $f = \frac{f_s}{K}k$, where f_s is the sampling frequency and K is the total number of frequency bins for each time frame.

$$X_{v,k} = Y_v \cdot (P_{nm,k} \circ (W \cdot D)) \tag{7}$$

where $\cdot$ denotes the matrix product operator and $\circ$ denotes the Hadamard (element-wise) matrix product, $X_{v,k} \in \mathbb{C}^{V \times K}$ contains all frequency STFT coefficients K of the signal for each virtual microphone v at the current time frame.

In (7), $P_{nm,k} \in \mathbb{C}^{(N+1)^2 \times K}$ contains the SH transform coefficients up to order N of the recorded pressure field. Following a general spatial sampling scheme [10], $P_{nm,k}$ can be discretely obtained at the current time frame from $X_q[n_f, k]$ as follows:

$$P_{nm,k} = Y^{\dagger} X_{q,k} \tag{8}$$

where the $\dagger$ operator denotes the pseudoinverse matrix, $X_{q,k} \in \mathbb{C}^{Q \times K}$ contains the the STFT coefficients for each microphone q and each frequency bin k of the current time frame. Specifically, let the row vector $x_q[k] = [X_q[n_f, 0] \cdots X_q[n_f, K-1]]$ represent the K STFT coefficients of microphone q at time frame n_f, the matrix $X_{q,k} = [x_0[k] \cdots x_{Q-1}[k]]^T$ is obtained by vertically stacking the row vectors $x_q[k]$ for each microphone q.

In a similar fashion, the matrix $Y \in \mathbb{C}^{Q \times (N+1)^2}$ presented in (8) contains the values of the spherical harmonics evaluated at each microphone location (θ_q, ϕ_q) up to the maximum order N. Specifically, let the row vector $Y_q = \left[Y_0^0(\theta_q, \phi_q), Y_1^{-1}(\theta_q, \phi_q) \cdots Y_N^N(\theta_q, \phi_q)\right]$ represent the $(N+1)^2$ spherical harmonics evaluated at the coordinates of microphone q. Then, the matrix $Y = [Y_0 \cdots Y_{Q-1}]^T$ is obtained by vertically stacking the row vectors Y_q for each microphone q.

The matrix $Y_v \in \mathbb{C}^{V \times (N+1)^2}$ presented in (7) is defined analogously to the matrix Y, containing the values of the spherical harmonics evaluated at each pointing direction $v \in \{0, 1, ..., V-1\}$ up to the maximum order N. This enables the steering of the beam pattern for each virtual microphone in the beamforming process.

At low frequencies, the function $|b_n(\kappa r)|$ tends to have small values for higher orders n, leading to significant amplification when computing the magnitude of $\frac{p_{nm}(\kappa, r)}{b_n(\kappa r)}$. As illustrated in Figure 1, this amplification can become problematic in practical implementations, where real spherical arrays introduce noise from the microphones. Excessive amplification in these conditions leads to a severe degradation of SNR, making it necessary to limit the gain applied to the pressure field coefficients.

To control this effect, the beamformer incorporates a crossover strategy in which the order n is incremented progressively at specific crossover frequencies. This ensures that amplification remains within a reasonable limit while preserving directional accuracy. In this implementation, the maximum allowed gain is empirically set to 40 dB (see Figure 1), providing a balance between maintaining spatial resolution and avoiding excessive noise amplification.

The matrix $W \in \mathbb{C}^{(N+1)^2 \times K}$ in (7) is designed to compensate for the amplitude of the SH transform coefficients of the pressure field. It consists of a vertical stack of rows, where each row corresponds to the multiplicative inverse of $b_n(\kappa r)$. However, to prevent excessive amplification at low frequencies, the corresponding entries in W for higher orders are replaced with zeros in these regions. This effectively enforces a crossover from lower orders to higher orders as frequency increases, as presented in Figure 1, ensuring that the beamformer maintains stability.

The matrix $D \in \mathbb{R}^{K \times K}$ in (7) is a diagonal matrix that contains the beamformer coefficients d_n associated with the maximum directivity beamformer, which corresponds to the hypercardioid polar pattern. These beamformer coefficients depend on the maximum order of the SH transform:

$$d_n = \frac{4\pi}{(N+1)^2}. \tag{9}$$

Since the maximum analysis order of the SH transform varies across frequency bins due to the crossover order scheme, the coefficients in D are filled accordingly. Specifically, D is a diagonal matrix where each diagonal element corresponds to a frequency bin k and follows the maximum analysis order at that frequency, resulting in $D = diag\left(\frac{4\pi}{(0+1)^2} \cdots \frac{4\pi}{(N+1)^2}\right)$.

After computing $X_{v,k}$, the next step is to reconstruct the time-domain representation of each virtual microphone signal v. This is achieved by applying the Inverse Short-Time Fourier Transform (ISTFT) algorithm, which converts the frequency domain data back into time-domain signals. Each virtual microphone's time-domain signal represents a spatially filtered version of the sound field, capturing the contribution from a specific direction or region of interest.

Since the STFT of the input signals has been modified, the ISTFT is not guaranteed to produce a realizable signal. To ensure proper reconstruction, the analysis and synthesis STFT windows must satisfy the Constant-Overlap Add (COLA) constraint, allowing the ISTFT to be computed using a least-squares estimation algorithm [13], [14].

B. Neural network model

The SH beamformer provides an initial source separation that contains signal bleed. To mitigate this, the proposed system incorporates a neural network-based de-bleeding stage, which refines the virtual microphone signals. This stage is implemented using Hybrid Demucs [15], a state-of-the-art deep learning model for source separation. Hybrid Demucs extends the Demucs architecture by incorporating both time-domain and time-frequency processing, leveraging the advantages of each representation to enhance signal separation.

Hybrid Demucs employs an encoder-decoder structure inspired by U-Net, where convolutional layers extract features from the input signals, and long short-term memory (LSTM) layers capture long-range temporal dependencies. The model operates on both raw waveforms and their spectrogram representations, fusing these complementary views to achieve high-fidelity separation. One of the key design choices in Hybrid Demucs is its hybrid processing approach, which allows it to maintain fine-grained temporal details while benefiting from spectrogram-based frequency resolution.

Hybrid Demucs was a baseline model in the Sound Demixing Challenge 2023 [4], highlighting its strong performance in separation tasks and making it a strong choice for de-bleeding tasks. In this work, we used the pretrained *hdemucs* model to this end, which can be retrieved using the official API[2].

IV. Results

To assess the separation performance of the SH beamformer, Hybrid Demucs and the combination of both, we consider three widely used objective metrics: Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), and Signal-to-Artifacts Ratio (SAR) [16]. The results are presented in Figure 3.

A. Experimental setup

To generate realistic test signals, the experiments were conducted using a Zylia ZM-1 spherical microphone array[3]. This array is capable of SH analysis up to order 3 and was deployed in a real-world recording environment. A set of RIRs were measured using the phase-controlled exponential sine sweep method presented in [17]. The measurement setup consisted of sound sources arranged in a circular configuration, evenly spaced at a distance of 1 meter from the array, within a room with a reverberation time (RT60) of approximately 328 ms. The obtained RIRs were then convolved with source signals from the MUSDB18-HQ dataset [18], generating test mixtures with 2 to 4 active sources.

As a commercial reference, we evaluated the Zylia Studio Pro software, which performs source separation by defining virtual microphones with adjustable directive patterns. To achieve the best possible separation performance, we configured the software to maximize directivity and to apply the highest available separation level, suppressing sound arriving from directions outside the defined virtual microphone beam pattern. We focused our evaluation on the most complex scenario, with four active sources, and did not assess mixtures with fewer sources due to the manual processing required for each test case.

For Hybrid Demucs, separation was performed on the signal from a single microphone at $(r_a, 0, 0)$ in spherical coordinates, positioned at the top of the array. This placement ensured similar energy levels across sources while minimizing attenuation from the rigid sphere's shadowing effect

[2] The pretrained model can be accessed through the API provided in the Demucs repository: https://github.com/facebookresearch/demucs/blob/main/demucs/pretrained.py.

[3] More information on the Zylia ZM-1: https://www.zylia.co/zylia-zm-1-microphone.html

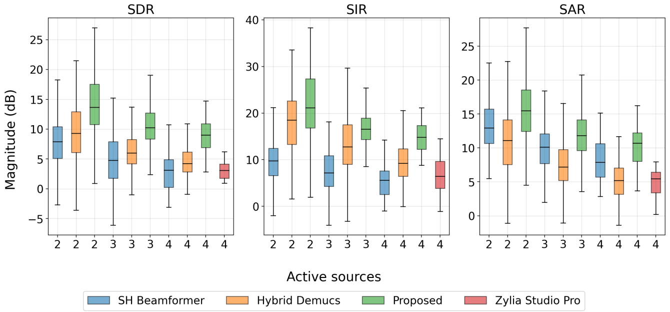

Fig. 3. Separation performance metrics for different numbers of active sound sources in a circular arrangement at 1 meter from the microphone array. The three plots show SDR, SIR, and SAR, measured in dB. Results correspond to the mean values across all active sources and are compared across four methods: the SH Beamformer (blue), Hybrid Demucs (orange), the Proposed method (green) and Zylia Studio Pro (red, only for 4 active sources).

For evaluation, reference signals corresponded to the isolated sources obtained after beamforming. In the case of Hybrid Demucs, the reference signals were the isolated signals captured by the same microphone used for separation.

B. Performance analysis

The results highlight the strengths and limitations of each method. The SH beamformer proves to be a powerful separation tool, achieving SDR values comparable to those obtained with the neural network model, particularly in scenarios with fewer active sound sources.

A key advantage of the SH beamformer is its ability to maintain higher SAR values, introducing minimal processing artifacts. In contrast, deep learning-based separation methods, while effective in reducing interference, often degrade SAR due to introduced distortions. This makes the SH beamformer a particularly suitable preprocessing stage for deep learning-based sound separation.

The Zylia Studio Pro software, evaluated in the 4-source scenario, exhibited slightly better SDR and SIR values compared to the SH beamformer. However, its separation performance was observed to be frequency-selective, leading to a degradation in SAR. This suggests that while Zylia's approach is effective in suppressing interference, it introduces artifacts that affect signal quality.

As depicted in Figure 3, our proposal consistently outperforms the other approaches, particularly in SIR, while maintaining stable SAR metrics, demonstrating its effectiveness in reducing interference while preserving signal quality. Although not real-time in this work, the system can be readily extended to low-latency operation as in Zylia Studio Pro.

V. Conclusions

In this work, we presented a spatial-informed, neural network-based de-bleeding approach that operates in the SH domain. Unlike previous studies that rely on simulated data, we evaluated our approach using real-world recordings captured with a spherical microphone array. The results highlight the synergistic relationship between the SH beamformer and deep learning-based source separation. The SH beamformer provides spatial filtering without introducing significant artifacts. By combining the SH beamformer with a neural network model, the resulting system enhances source separation performance while maintaining high signal quality. Moreover, we provide an open-source implementation of the described SH beamformer for efficient spatial filtering and virtual microphone synthesis, aiming to encourage further research and adoption of SH-domain processing in real-world audio scenarios.

References

[1] D. Arteaga, "Introduction to ambisonics," Online, may 2023. [Online]. Available: https://doi.org/10.5281/zenodo.7963106

[2] A. Farina, A. Capra, L. Chiesi, and L. Scopece, "A spherical microphone array for synthesizing virtual directive microphones in live broadcasting and in post production," in *Proceedings of the 40th International Conference of the Audio Engineering Society (AES)*, no. 3-1. Audio Engineering Society, October 2010.

[3] H. Sun, S. Yan, U. P. Svensson, and H.-F. Sun, "Spherical harmonics based optimal minimum sidelobe beamforming for spherical sensor arrays," in *2010 International ITG Workshop on Smart Antennas (WSA)*, 2010, pp. 286–291.

[4] G. Fabbro *et al.*, "The sound demixing challenge 2023 – music demixing track," *Transactions of the International Society for Music Information Retrieval*, vol. 7, no. 1, p. 63–84, 2024.

[5] Lluís, Francesc, Meyer-Kahlen, Nils, Chatziioannou, Vasileios, and Hofmann, Alex, "Direction specific ambisonics source separation with end-to-end deep learning," vol. 7, p. 29, 2023.

[6] M. Hafsati, N. Epain, R. Gribonval, and N. Bertin, "Sound source separation in the higher order ambisonics domain," in *Proceedings of the 22nd International Conference on Digital Audio Effects (DAFx-19)*, Birmingham, UK, September 2019.

[7] A. Herzog, S. R. Chetupalli, and E. A. P. Habets, "Ambisep: Ambisonic-to-ambisonic reverberant speech separation using transformer networks," 2022.

[8] M. Guzik and K. Kowalczyk, "On ambisonic source separation with spatially informed non-negative tensor factorization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3238–3255, 2024.

[9] Rafaely, Boaz, "Spatial alignment of acoustic sources based on spherical harmonics radiation analysis," in *2010 4th International Symposium on Communications, Control and Signal Processing (ISCCSP)*, 2010, pp. 1–5.

[10] B. Rafaely, *Fundamentals of Spherical Array Processing*, 2nd ed., ser. Springer Topics in Signal Processing. Springer, Cham, 2019, vol. 9.

[11] J. Meyer and G. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 2002, pp. II–1781–II–1784.

[12] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *The time-dependent Fourier transform*, 2nd ed. Prentice Hall, 1999, ch. 10.3, pp. 714–722.

[13] B. Sharpe, "Invertibility of overlap-add processing," Online, accessed March 2025. [Online]. Available: https://gauss256.github.io/blog/cola.html

[14] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[15] A. Défossez, "Hybrid spectrogram and waveform source separation," in *Proceedings of the ISMIR 2021 Workshop on Music Source Separation*, 2021.

[16] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[17] K. Vetter and S. di Rosario, "Expochirptoolbox: a pure data implementation of ess impulse response measurement," in *4th Pure Data Convention*, 2011.

[18] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "MUSDB18-HQ - an uncompressed version of musdb18," Dec. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3338373